

# On approximations via convolution-defined mixture models

Hien D. Nguyen\* and Geoffrey J. McLachlan\*

December 29, 2016

## Abstract

An often-cited fact regarding mixing distributions is that their densities can approximate the densities of any unknown distribution to arbitrary degrees of accuracy provided that the mixing distribution is sufficiently complex. This fact is often not made concrete. We investigate theorems that provide approximation bounds for mixing distributions. Novel connections are drawn between the approximation bounds of mixing distributions and estimation bounds for the maximum likelihood estimator of finite mixtures of location-scale distributions. New approximation and estimation bounds are obtained in the context of finite mixtures of truncated location-scale distributions.

## 1 Introduction

Mixture models are an important class of probability models that have found use in all areas of application in statistics, machine learning, and beyond. An important class of mixture models are the mixing distributions with probability density functions (PDFs) of the form  $f(\mathbf{x}) = \int_{\mathbb{X}} f(\mathbf{x}; \boldsymbol{\theta}) d\Pi(\boldsymbol{\theta})$ , where  $f(\cdot; \boldsymbol{\theta})$  is a PDF that is dependent on some parameter  $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$  with distribution function (DF)  $\Pi(\boldsymbol{\theta})$ . Here  $\mathbb{X} \subseteq \mathbb{R}^p$  is the support of the PDFs  $f$  and  $f(\cdot; \boldsymbol{\theta})$ . Notice that the mixing distributions contain the finite mixture models by setting the DF to  $\Pi(\boldsymbol{\theta}) = \sum_{i=1}^n \pi_i \delta(\mathbf{x} - \boldsymbol{\theta}_i)$ , for  $n \in \mathbb{N}$ , where  $\delta$  is the Dirac delta function and  $\mathbb{N}$  is the natural numbers; see

---

\*Hien Nguyen and Geoffrey McLachlan are with the School of Mathematics and Physics, The University of Queensland, St. Lucia, Brisbane, Australia 4075. Hien Nguyen is with the Centre for Advanced Imaging, The University of Queensland, St. Lucia, Brisbane, Australia 4075. \*Corresponding Author: Hien Nguyen (Email: h.nguyen7@uq.edu.au).

Lindsay (1995, Sec. 1.2.2) and McLachlan & Peel (2000, Sec. 1.12) for descriptions and references regarding mixing distributions; Leroux (2006) provides a more recent reference on the topic.

The appeal of mixture models largely comes from their flexibility of representation. The folk theorem regarding mixture models generally states that a mixture model can approximate any distribution to a sufficient level of accuracy, provided that the mixing distribution of the mixture model is sufficiently complex. Example statements of the folk theorem include: “there is an obvious sense in which the mixture of normals approach, given enough components, can approximate any multivariate density ” (Rossi, 2014, p. 5), “the (mixture) model forms can fit any distribution and significantly increase model fit ” (Walker & Ben-Akiva, 2011, p. 173), “a mixture model can approximate almost any distribution” (Yona, 2011, p. 500). From the examples, we see that statements regarding the flexibility of mixture models are generally left vague and unclear.

Let  $d_{\mathbb{X}}^{\text{TV}}(f, g) = 2^{-1} \|f - g\|_{\mathbb{X},1}$  be the total-variation distance, where  $\|f\|_{\mathbb{X},q} = \left(\int_{\mathbb{X}} |f(\mathbf{x})|^q d\mathbf{x}\right)^{1/q}$  for  $1 \leq q < \infty$  is the  $\mathcal{L}_q$ -norm. Here,  $f$  and  $g$  are functions over the set  $\mathbb{X} \subseteq \mathbb{R}^d$ . We also let  $\|f\|_{\mathbb{X},\infty} = \sup_{\mathbf{x} \in \mathbb{X}} |f(\mathbf{x})|$ . Further define the location-scale family of PDFs over  $\mathbb{R}$  as

$$\mathcal{F}^1 = \left\{ f : \int_{\mathbb{R}} \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) dx = 1, \text{ for all } \mu \in \mathbb{R} \text{ and } \sigma \in \mathbb{R}^+ \right\},$$

where  $\mathbb{R}_+ = (0, \infty)$ , and let  $x_i$  refer to the  $i$ th element of  $\mathbf{x}$ , for  $i \in [p]$  ( $[p] = 1, \dots, p$ ). In DasGupta (2008, Sec. 33.1), the following theorem is stated.

**Theorem 1** (DasGupta 2008, Thm. 33.1). *Let  $f$  be a probability density function over  $\mathbb{R}^p$  for  $p \in \mathbb{N}$ . If  $\mathcal{F}_g^2$  is the class of mixtures of  $g \in \mathcal{F}^1$ :*

$$\mathcal{F}_g^2 = \left\{ f^* : f^*(\mathbf{x}) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^p} \frac{1}{\sigma^d} \prod_{i=1}^p g\left(\frac{x_i - \mu_i}{\sigma}\right) d\Pi_{\mu}(\boldsymbol{\mu}) d\Pi_{\sigma}(\sigma) \right\}$$

*Then, given any  $\epsilon > 0$ , there exists a  $f^* \in \mathcal{F}^2$  such that  $d_{\mathbb{R}^d}^{\text{TV}}(f, f^*) < \epsilon$ , where  $\Pi_{\mu}$  and  $\Pi_{\sigma}$  are DFs over  $\mathbb{R}^d$  and  $\mathbb{R}_+$ , respectively.*

Upon inspection, Theorem 1 states that the class of marginally-independent location-scale mixing distributions can approximate distributions with any PDFs arbitrarily well, with respect to the total-variation distance. Unfortunately, the proof of Theorem 1 is not provided in DasGupta (2008). Remarks regarding the theorem defer the proof to unknown locations in Cheney & Light (2000), which makes it difficult to investigate the structure and nature of the theorem.

In this article, we investigate the proofs from Cheney & Light (2000) and consider alternative versions of Theorem 1 that provide more insight into the structures of the results. For example, we demonstrate that a weaker alternative to Theorem 1 is possible, whereupon only a mixture over the location parameter is required. That is, no integration over the scale parameter element is needed, as in  $\mathcal{F}_g^2$ . Furthermore, we show that stronger results than total variation approximation is possible in Theorem 1. In particular, within compact subsets of the support of the target PDF, uniform approximation is possible. Rates of convergence are also possible if we make a Lipschitz assumption on the target PDF.

In addition to the presentation of Theorem 1 and its variants, we also consider the relationship between the mixing distributions results and the finite mixture model estimation approximation bounds of Zeevi & Meir (1997). Via the approximation and estimation bounding results for the maximum likelihood estimator (MLE) from Li & Barron (1999) and Rakhlin et al. (2005), we further present new results for bounding Kullback-Leibler errors [KL; (Kullback & Leibler, 1951)] for the MLE of finite mixtures of location-scale PDFs. Finally, we consider the problem of approximating compactly supported PDFs via mixtures of truncated location-scale PDFs over the same domain.

We proceed as follows. In Section 2, we discuss Theorem 1 and its variants. The relationship between the mixing distributions approximation results and the results of Zeevi & Meir (1997), Li & Barron (1999), and Rakhlin et al. (2005) are then presented in Sections 3, 4, and 5, respectively. Truncated approximations are finally considered in Section 6.

## 2 Mixing distributions approximation theorems

Let  $\mathcal{L}_q(\mathbb{X})$  be the space of functions having the property  $\|f\|_{\mathbb{X},q} < \infty$ , with support  $\mathbb{X}$ , and which map to  $\mathbb{R}$ . Further, we define the convolution between  $f \in \mathcal{L}_q(\mathbb{X})$  and  $g \in \mathcal{L}_r(\mathbb{X})$  as

$$(f * g)(\mathbf{x}) = \int_{\mathbb{X}} f(\mathbf{y}) g(\mathbf{x} - \mathbf{y}) d\mathbf{y}, \quad (1)$$

where (1) exists and is measurable for specific cases, due to results such as the following from Makarov & Podkorytov (2013, Sec. 9.3).

**Theorem 2** (Makarov and Podkorytov, 2013, Sec. 9.3.1-2). *Let  $f \in \mathcal{L}_q(\mathbb{R}^p)$  and  $g \in \mathcal{L}_r(\mathbb{R}^p)$ , for  $q, r \in \mathbb{N} \cup \{\infty\}$ .*

We have the following results:

- (i) if  $q = 1$ , then  $f * g$  exists and  $\|f * g\|_{\mathbb{R}^p, r} \leq \|f\|_{\mathbb{R}^p, 1} \|g\|_{\mathbb{R}^p, r}$ .
- (ii) if  $1/q + 1/r = 1$ , then  $f * g$  exists and  $\|f * g\|_{\mathbb{R}^p, \infty} \leq \|f\|_{\mathbb{R}^p, q} \|g\|_{\mathbb{R}^p, r}$ .

*Remark 3.* When  $q = r = 1$ , not only do we have inequality (i) of Theorem 2, but also that

$$\begin{aligned} \int_{\mathbb{R}^p} (f * g)(\mathbf{x}) d\mathbf{x} &= \int_{\mathbb{R}^p} f(\mathbf{y}) \int_{\mathbb{R}^p} g(\mathbf{x} - \mathbf{y}) d\mathbf{x} d\mathbf{y} \\ &= \int_{\mathbb{R}^p} f(\mathbf{x}) d\mathbf{x} \int_{\mathbb{R}^p} g(\mathbf{x}) d\mathbf{x} < \infty. \end{aligned}$$

Moreover, this implies that if  $f$  and  $g$  are PDFs over  $\mathbb{R}^p$ , then  $f * g$  is also a PDF over  $\mathbb{R}^p$ .

Let a function  $\alpha_k \in \mathcal{L}_1(\mathbb{R}^p)$  for  $k \in \mathbb{R}^+$  be called an approximate identity in  $\mathbb{R}^p$  if there exists a  $k^* \in \mathbb{R}^+ \cup \{0, \infty\}$  such that (i)  $\alpha_k \geq 0$ , (ii)  $\int_{\mathbb{R}^p} \alpha_k(\mathbf{x}) d\mathbf{x} = 1$ , and (iii)  $\int_{\|\mathbf{x}\|_1 > \delta} \alpha_k(\mathbf{x}) d\mathbf{x} \rightarrow 0$  as  $k \rightarrow k^*$ , for every  $\delta > 0$  [cf. Makarov & Podkorytov (2013, Sec. 7.6.1)]. Here,  $\|\mathbf{x}\|_q = (\sum_{i=1}^p |x_i|^q)^{1/q}$  is the  $l_q$ -vector norm. The following result of Cheney & Light (2000) provides a useful generative method for constructing approximate identities.

**Lemma 4** (Cheney and Light, 2000, Ch. 20, Thm. 4). *Let  $\alpha \in \mathcal{L}_1(\mathbb{R}^p)$  and let  $k \in \mathbb{N}$ . If  $\alpha \in \mathcal{F}^3$ , where*

$$\mathcal{F}^3 = \left\{ f \in \mathcal{L}_1(\mathbb{R}^p) : f(\mathbf{x}) = \prod_{i=1}^p g(x_i), g \in \mathcal{F}^1 \right\},$$

*then the dilations  $\alpha_k(\mathbf{x}) = k^p \alpha(k\mathbf{x})$  is an approximate identity, for  $k^* = \infty$ .*

As alluded to in the introduction, we may call  $\mathcal{F}^3$  the class of marginally-independent location-scale density functions. With an ability to construct approximate identities, the following theorem from Makarov & Podkorytov (2013) provides a powerful means to construct approximations for any function over  $\mathbb{R}^p$ . The corollary to the result provides a statistical interpretation.

**Theorem 5** (Makarov and Podkorytov, 2013, Sec. 9.3.3). *Let  $\alpha_k$  be an approximate identity in  $\mathbb{R}^p$  for some  $k^* \in \mathbb{R}^+ \cup \{0, \infty\}$ . If  $f \in \mathcal{L}_q(\mathbb{R}^p)$  for  $q \in \mathbb{N}$ , then  $\|f * \alpha_k - f\|_{\mathbb{R}^p, q} \rightarrow 0$  as  $k \rightarrow k^*$ .*

**Corollary 6.** *Let  $f$  be a PDF over  $\mathbb{R}^d$ . If  $g \in \mathcal{F}^3$ , then for any  $\epsilon > 0$ , there exists a PDF  $f^*$  in*

$$\mathcal{F}_g^4 = \left\{ f^* : f^*(\mathbf{x}) = \int_{\mathbb{R}^p} k^p g(k\mathbf{x} - \mathbf{m}) f(\mathbf{m}) d\mathbf{m}, k \in \mathbb{N} \right\}$$

*such that  $\|f - f^*\|_{\mathbb{R}^p, q} < \epsilon$ , for any  $q \in \mathbb{R}^+$ .*

*Proof.* From Lemma 4 and Theorem 5, for any  $g \in \mathcal{F}^3$  and  $f \in \mathcal{L}_q(\mathbb{R}^p)$ , we have  $\|f * [k^p g(k \times \cdot)] - f\|_{\mathbb{R}^p, q} \rightarrow 0$ , where the convolution  $f * [k^p g(k \times \cdot)] = \int_{\mathbb{R}^p} k^p g(k\mathbf{x} - \mathbf{m}) f(\mathbf{m}) d\mathbf{m}$ . By the definition of convergence, we have the fact that for every  $\epsilon > 0$ , there exists some  $K$  such that for all  $k > K$ ,  $\|f * [k^p g(k \times \cdot)] - f\|_{\mathbb{R}^p, q} < \epsilon$ . Putting the convolutions  $f * [k^p g(k \times \cdot)]$  for all  $k \in \mathbb{N}$  into  $\mathcal{F}_g^4$  provides the desired convergence result by noting that  $\mathcal{L}_1(\mathbb{R}^p) \subset \mathcal{L}_q(\mathbb{R}^p)$  when  $q > 1$ . Lastly,  $\bar{f}$  is a PDF via Remark 3.  $\square$

*Remark 7.* Corollary 6 improves upon Theorem 1 in several ways. Firstly, the total variation bound is replaced by the stronger  $\mathcal{L}_q$ -norm result. Secondly, mixing only occurs over the mean parameter element  $\mathbf{m}$ , via the PDF  $d\Pi_m(\mathbf{m})/d\mathbf{m} = f(\mathbf{m})$ , and not over the scaling parameter element  $k$ , which can be taken as a constant value. That is, we only require that the class  $\mathcal{F}_g^4$  be mixing distributions over the location and scale parameter elements of  $g$ , where  $\Pi_m$  is the DF that is determined by the density being approximated and the scale parameter element is picked to be some fixed value  $k \in \mathbb{N}$ . Lastly, we note that Theorem 1 can simply be obtained as the  $q = 1$  case of Corollary 6 by setting  $\sigma = 1/k$  and  $\boldsymbol{\mu} = \mathbf{m}/k$ .

Notice that Theorem 5 cannot be used to provide  $\mathcal{L}_\infty$ -norm approximation results. Let  $\mathcal{C}(\mathbb{X})$  be the class of continuous functions over the set  $\mathbb{X}$ . If one assumes that the target PDF  $f$  is bounded and belongs to  $\mathcal{C}(\mathbb{R}^p)$ , then a uniform approximation alternative to Theorem 5 is possible for compact subsets of  $\mathbb{R}^p$ .

**Theorem 8** (Cheney and Light, 2000, Ch. 20, Thm. 2). *Let  $\alpha_k$  be an approximate identity in  $\mathbb{R}^p$  for some  $k^* \in \mathbb{R}^+ \cup \{0, \infty\}$ . If  $f$  is a bounded function in  $\mathcal{C}(\mathbb{R}^p)$ , then  $\|f * \alpha_k - f\|_{\mathbb{K}, \infty} \rightarrow 0$  as  $k \rightarrow k^*$ , for all compact  $\mathbb{K} \subset \mathbb{R}^p$ .*

We note in passing that Theorem 8 can be used to prove density results for finite mixture models, such as that of DasGupta (2008, Thm. 33.2). For further details, see Cheney & Light (2000, Thm. 5) and Light (1993). Let  $\text{Lip}_a(\mathbb{X})$  be the class of Lipschitz functions  $f$ , where  $|f(\mathbf{x}) - f(\mathbf{y})| \leq C \|\mathbf{x} - \mathbf{y}\|_\infty^a$ , for some  $a, C \in \mathbb{R}^+$ , where

$\mathbf{x}, \mathbf{y} \in \mathbb{X}$ . If one assumes that the target PDF is in  $\text{Lip}_a(\mathbb{X})$  for some  $a \in (0, 1]$ , then the following approximation rate result is available.

**Theorem 9** (Cheney and Light, 2000, Ch. 21, Thm. 1). *Let  $\alpha_k$  be an approximate identity in  $\mathbb{R}^p$  for some  $k^* \in \mathbb{R}^+ \cup \{0, \infty\}$  with the additional property that  $\int_{\mathbb{R}^p} \|\mathbf{x}\|_1^a \alpha(\mathbf{x}) d\mathbf{x} < \infty$  for some  $a \in (0, 1]$ . If  $f \in \text{Lip}_a(\mathbb{R}^p)$ , then there exists a constant  $A > 0$  such that  $\|f * \alpha_k - f\|_{\mathbb{R}^p, \infty} \leq A/k^a$  for  $k \in \mathbb{N}$ .*

**Example 10.** Let  $\alpha \in \mathcal{F}^3$  be generated by taking the marginal location-scale density  $g = \phi$ , where  $\phi \in \mathcal{F}^1$  is the standard normal PDF. The condition  $\int_{\mathbb{R}^p} \|\mathbf{x}\|_1^a \alpha(\mathbf{x}) d\mathbf{x} < \infty$  is satisfied for  $a = 1$  since the multivariate normal distribution has all of its polynomial moments; see for example Willink (2005).

**Corollary 11.** *Let  $f \in \text{Lip}_1(\mathbb{R}^d)$  be a PDF. If  $f^*(\mathbf{x}) = \int_{\mathbb{R}^d} k^p \prod_{i=1}^p \phi(kx_i - m_i) f(\mathbf{m}) d\mathbf{m}$ , then  $\|f - f^*\|_{\mathbb{R}^p, \infty} \leq A/k$  for  $k \in \mathbb{N}$  and some constant  $A > 0$ .*

Thus, the mixing distribution generated via marginally-independent normal PDFs converges uniformly for target PDFs  $f \in \text{Lip}_1(\mathbb{R}^p)$ , at a rate of  $1/k$ .

### 3 Bounding of Kullback-Leibler divergence via results from Zeevi and Meir (1997)

Let  $\mathbb{K} \subset \mathbb{R}^p$  be a compact subset and let

$$\mathcal{F}_{\mathbb{K}, \beta}^5 = \left\{ f : \int_{\mathbb{K}} f(\mathbf{x}) d\mathbf{x} \text{ and } f(\mathbf{x}) \geq \beta > 0 \text{ for all } \mathbf{x} \in \mathbb{K} \right\}$$

be the class of lower-bounded target PDFs over  $\mathbb{K}$ . In Zeevi & Meir (1997), the approximation errors for finite mixtures of marginally-independent symmetric location-scale PDFs are studied in the context of approximating functions in  $\mathcal{F}_{\mathbb{K}, \beta}^5$ .

*Remark 12.* The use of finite mixtures of marginally-independent location-scale PDFs is implicit in Zeevi & Meir (1997) as they report on approximation via product kernels of radial basis functions. The products of kernels is equivalent to taking products over marginally-independent densities to yield a joint density. Univariate radial basis functions that are positive and integrate to unity are symmetric location-scale PDFs in one dimension. Thus, the

product of univariate radial basis functions that generate PDFs correspond to a subclass of  $\mathcal{F}^3$ ; see Buhmann (2003) regarding radial basis functions.

Let the KL divergence between two PDFs  $f, g \in \mathcal{L}_1(\mathbb{X})$  be defined as

$$d_{\mathbb{X}}^{\text{KL}}(f, g) = \int_{\mathbb{X}} f(\mathbf{x}) \log \left[ \frac{f(\mathbf{x})}{g(\mathbf{x})} \right] d\mathbf{x}.$$

The KL divergence between  $f$  and  $g$  is difficult to work with as it is not a distance function, it is asymmetric, and it does not obey the triangle inequality. As such, bounding the KL divergence by a distance function provides a useful means of manipulation and control. The following useful result is obtained by Zeevi & Meir (1997).

**Lemma 13** (Zeevi and Meir, 1997, Lemma 3). *If  $f, g \in \mathcal{F}_{\mathbb{K}, \beta}^5$ , then  $d_{\mathbb{K}}^{\text{KL}}(f, g) \leq \beta^{-1} \|f - g\|_{\mathbb{K}, 2}^2$ .*

For any  $g \in \mathcal{F}^3$ , define the  $n$ -component bounded finite mixtures of  $g$  as the class

$$\mathcal{F}_{g, n}^6 = \left\{ f : f(\mathbf{x}) = \sum_{i=1}^n \pi_i k_i^p g(k_i \mathbf{x} - \mathbf{m}_i), \right. \\ \left. \mathbf{m}_i \in [\underline{m}, \overline{m}]^p, k_i \in [\underline{k}, \overline{k}], \pi_i \geq 0, \text{ and } \sum_{i=1}^n \pi_i = 1 \right\},$$

where  $i \in [n]$ . Here,  $-\infty < \underline{m} < \overline{m} < \infty$  and  $0 < \underline{k} < \overline{k} < \infty$  ensure the boundedness of functions of  $\mathcal{F}_{g, n}^6$ .

Let  $\mathcal{F}_g^7 = \{k^p g(k\mathbf{x} - \mathbf{m}) : \mathbf{m} \in [\underline{m}, \overline{m}]^p \text{ and } k \in [\underline{k}, \overline{k}]\}$ , and let  $\text{Conv}_n(\mathcal{F}_g^7) = \mathcal{F}_{g, n}^6$  denote the  $n$ -point convex hull of  $\mathcal{F}_g^7$ . We simply refer to  $\text{Conv}_{\infty}(\mathcal{F}_g^7) = \text{Conv}(\mathcal{F}_g^7)$  as the convex hull. The closure of  $\text{Conv}(\mathcal{F}_g^7)$  can then be defined as

$$\overline{\text{Conv}}(\mathcal{F}_g^7) = \left\{ f : f(\mathbf{x}) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^p} k^p g(k\mathbf{x} - \mathbf{m}) d\Pi_m d\Pi_k, \right. \\ \left. \Pi_m \in M_m, \text{ and } \Pi_k \in M_k \right\}$$

where  $M_m$  and  $M_k$  are the sets of all probability measures over  $\mathbf{m}$  and  $k$ , respectively. See van de Geer (2003) for clear definitions of convex hulls and their closures. The following result from Barron (1993) relates the closure of convex hulls to the  $\mathcal{L}_2$ -norm.

**Lemma 14** (Barron, 1993, Lemma 1). *If  $\bar{f}$  is in  $\overline{\text{Conv}}(\mathcal{F})$ , where  $\mathcal{F}$  is in a Hilbert space over support  $\mathbb{X}$  such that  $\|f\|_{\mathbb{X}, 2}^2 \leq B$  for each  $f \in \mathcal{F}$ , then for every  $n \in \mathbb{N}$ , and every  $C > B^2 - \|\bar{f}\|_{\mathbb{X}, 2}^2$ , there exists an  $f_n \in \text{Conv}_n(\mathcal{F})$  such that  $\|\bar{f} - f_n\|_{\mathbb{X}, 2}^2 \leq C/n$ .*

Thus, from Lemma 14, we know that if  $\bar{f} \in \overline{\text{Conv}}(\mathcal{F}_g^7)$ , then there exists an  $n$ -component finite mixture of density  $g$ ,  $f_n \in \text{Conv}_n(\mathcal{F}_g^7)$  such that  $\|\bar{f} - f_n\|_{\mathbb{K},2}^2 \leq C/n$ , where  $\mathbb{K}$  is the compact support of both densities and  $C > 0$  is a constant that depends on the class  $\mathcal{F}_g^7$ , which we know to be bounded. From Corollary 6 we know that if  $f \in \mathcal{F}_{\mathbb{K},\beta}^5 \cap \mathcal{L}_2(\mathbb{K})$ , then for every  $\epsilon > 0$ , there exists an  $\bar{f} \in \mathcal{F}_g^4$  such that for an  $\|\bar{f} - f\|_{\mathbb{K},2}^2 < \epsilon$ . Since  $\mathcal{F}_g^4 \subset \overline{\text{Conv}}(\mathcal{F}_g^7)$ , we can set  $\bar{f} = f^*$ . An application of the triangle inequality yields the following result from Zeevi & Meir (1997).

**Theorem 15** (Zeevi and Meir, 1997, Eqn. 27). *If  $f \in \mathcal{F}_{\mathbb{K},\beta}^5 \cap \mathcal{L}_2(\mathbb{K})$ , then for any  $\epsilon > 0$  and  $g \in \mathcal{F}^3$ , there exists an  $f_n \in \text{Conv}_n(\mathcal{F}_g^7)$  such that  $d_{\mathbb{K}}^{KL}(f, f_n) \leq \epsilon/\beta + C/(k\beta)$ , for some  $C > 0$  and  $n \in \mathbb{N}$ .*

*Proof.* By the triangle inequality, we have  $\|f_n - f\|_{\mathbb{K},2}^2 \leq \|f_n - \bar{f}\|_{\mathbb{K},2}^2 + \|\bar{f} - f\|_{\mathbb{K},2}^2 \leq \epsilon + C/n$ . We then apply Lemma 13 to obtain the desired result.  $\square$

*Remark 16.* The application of Corollary 6 requires the convolution of a compactly supported function with a function over  $\mathbb{R}^p$ . In general, the convolution of two functions on different supports produces a function with a support that is a function of the original supports. That is, if  $f$  is supported on  $\text{supp}(f)$  and  $g$  is supported on  $\text{supp}(g)$ , then the support of  $f * g$  is a subset of the closure of the set  $\{\mathbf{x} + \mathbf{y} : \mathbf{x} \in \text{supp}(f), \mathbf{y} \in \text{supp}(g)\}$ . In order to mitigate against the calculus of supports, we can allow any compactly supported PDF  $f$  to take values outside of its support  $\mathbb{K}$  by simply setting  $f(\mathbf{x}) = 0$  if  $\mathbf{x} \notin \mathbb{K}$  and thus implicitly only work with functions over  $\mathbb{R}^p$ .

*Remark 17.* We note that Zeevi & Meir (1997) utilized a slightly different version of Corollary 6 that makes use of the alternative approximate identity  $\alpha_k(\mathbf{x}) = k^{-p}g(\mathbf{x}/k)$  with  $k^* = 0$ . Here,  $g$  is taken to be a product kernel of radial basis functions.

An approach for quantifying the error of the quasi-maximum likelihood estimator (quasi-MLE) for finite mixture models, with respect to the Hellinger divergence, is then developed by Zeevi & Meir (1997) via the theory of White (1982). We will instead pursue the bounding of KL errors for the MLE via the directions of Li & Barron (1999) and Rakhlin et al. (2005).



## 4 Maximum likelihood estimation bounds via results from Li and Barron (1999)

As alternatives to Lemma 14 and Theorem 15, we can interpret the following two results from Li & Barron (1999) for finite mixtures of location-scale PDFs over compact supports  $\mathbb{K}$ .

**Theorem 18** (Li and Barron, 1999, Thm. 1). *If  $g \in \mathcal{F}^3$  and  $\bar{f} \in \overline{\text{Conv}}(\mathcal{F}_g^7)$ , then there exists an  $f_n \in \text{Conv}_n(\mathcal{F}_g^7)$  such that  $d_{\mathbb{K}}^{KL}(f, f_n) \leq C\gamma/n$ , where*

$$C = \int_{\mathbb{K}} \frac{\int_{\mathbb{K}} [k^p g(k\mathbf{x} - \mathbf{m})]^2 d\Pi_m d\Pi_k}{\int_{\mathbb{K}} k^p g(k\mathbf{x} - \mathbf{m}) d\Pi_m d\Pi_k} d\mathbf{x}$$

with DFs  $\Pi_m$  and  $\Pi_k$  over  $\mathbb{R}^p$  and  $\mathbb{R}^+$ , respectively, and  $\gamma = 4(\log(3\sqrt{e}) + A)$  with

$$A = \sup_{\mathbf{m}_1, \mathbf{m}_2, k_1, k_2, \mathbf{x}} \log \frac{k_1^p g(k_1 \mathbf{x} - \mathbf{m}_1)}{k_2^p g(k_2 \mathbf{x} - \mathbf{m}_2)}.$$

**Theorem 19** (Li and Barron, 1999, Thm. 2). *If  $f \in \mathcal{F}_{\mathbb{K}, \beta}^5$  and  $g \in \mathcal{F}^3$ , then there exists an  $f_n \in \text{Conv}_n(\mathcal{F}_g^7)$  such that  $d_{\mathbb{K}}^{KL}(f, f_n) \leq d_{\mathbb{K}}^{KL}(f, \bar{f}^*) + C^*\gamma/n$ , where  $\gamma$  is as defined in Theorem 18,*

$$\bar{f}^* = \inf_{\bar{f} \in \overline{\text{Conv}}(\mathcal{F}_g^7)} d_{\mathbb{K}}^{KL}(f, \bar{f})$$

and  $C^*$  is the smallest limit of

$$C = \int_{\mathbb{K}} \frac{\int_{\mathbb{K}} [k^p g(k\mathbf{x} - \mathbf{m})]^2 d\Pi_m d\Pi_k}{\left(\int_{\mathbb{K}} k^p g(k\mathbf{x} - \mathbf{m}) d\Pi_m d\Pi_k\right)^2} f(\mathbf{x}) d\mathbf{x},$$

for sequences of  $\Pi_m$  and  $\Pi_k$  achieving  $d_{\mathbb{K}}^{KL}(f, \bar{f})$  values that approach  $d_{\mathbb{K}}^{KL}(f, \bar{f}^*)$ .

*Remark 20.* Upon first inspection, it may appear as if there is no need to restrict the domain of application of Theorems 18 and 19 to compact supports. However, if one notes the constant  $A$  in Theorem 18, it is clear that for any location-scale PDFs over  $\mathbb{R}^p$ , if  $\|\mathbf{x}\|_1 \rightarrow \infty$ , then the denominator goes to zero and thus the expression is undefined. Thus, one often needs to restrict both the domain of the parameter elements and the support of the

PDFs in order to define  $A$ .

**Corollary 21.** *If  $f \in \mathcal{F}_{\mathbb{K},\beta}^5$  and  $g \in \mathcal{F}^3$ , then for any  $\epsilon > 0$ , there exists an  $f_n \in \text{Conv}_n(\mathcal{F}_g^7)$  such that  $d_{\mathbb{K}}^{\text{KL}}(f, f_n) \leq \epsilon/\beta + C^*\gamma/n$ , where  $\gamma$  and  $C^*$  are as defined in Theorems 18 and 19.*

*Proof.* By definition, for any  $\bar{f} \in \overline{\text{Conv}}(\mathcal{F}_g^7)$ ,  $d_{\mathbb{K}}^{\text{KL}}(f, \bar{f}^*) \leq d_{\mathbb{K}}^{\text{KL}}(f, \bar{f})$ . However, by Corollary 6 and Lemma 13, for every  $\epsilon > 0$ , there exists an  $\bar{f}$  such that  $d_{\mathbb{K}}^{\text{KL}}(f, \bar{f}) < \epsilon/\beta$ .  $\square$

Corollary 21 implies that can approximate a compactly supported PDF to arbitrary degrees of accuracy using finite mixtures of location-scale PDFs of increasing numbers of components  $n$ . Thus far, the results have focused on functional approximation, and not bounding the estimators from data. We now present a KL error bounding result for the MLE.

Let  $\mathbf{X}_1, \dots, \mathbf{X}_N$  be  $N$  independent and identically distributed (IID) random sample generated from a distribution with density  $f \in \mathcal{F}_{\mathbb{K},\beta}^5$ . Define the log-likelihood function of an  $n$ -component mixture of location-scale PDFs  $g \in \mathcal{F}^3$  from the realizations  $\mathbf{x}_1, \dots, \mathbf{x}_N$  as

$$\ell_{g,n,N}(\boldsymbol{\theta}) = \sum_{j=1}^N \log \left[ \sum_{i=1}^n \pi_i k_i^p g(k_i \mathbf{x}_i - \mathbf{m}_i) \right],$$

where  $\boldsymbol{\theta}$  contains  $\pi_i$ ,  $k_i$ , and  $\mathbf{m}_i$  for  $i \in [n]$ . The MLE can then be defined as

$$\hat{f}_{g,n,N}(\mathbf{x}) = \sum_{i=1}^n \hat{\pi}_i \hat{k}_i^p g(\hat{k}_i \mathbf{x} - \hat{\mathbf{m}}_i),$$

where

$$\hat{\boldsymbol{\theta}}_{n,N} \in \left\{ \hat{\boldsymbol{\theta}} : \ell_{g,n,N}(\hat{\boldsymbol{\theta}}) = \max \ell_{g,n,N}(\boldsymbol{\theta}), \right. \\ \left. \text{satisfying the restrictions of } \mathcal{F}_g^7 \right\}$$

The elements  $\hat{\pi}_i$ ,  $\hat{k}_i$ , and  $\hat{\mathbf{m}}_i$ , in  $\hat{\boldsymbol{\theta}}_{n,N}$ , are the corresponding elements of  $\pi_i$ ,  $k_i$ , and  $\mathbf{m}_i$ .

For  $g \in \mathcal{F}^3$ , consider the family of location PDFs  $\mathcal{F}_{g,k}^8 = \{k^p g(k\mathbf{x} - \mathbf{m}) : \mathbf{m} \in [\underline{\mathbf{m}}, \overline{\mathbf{m}}]^p\}$ , where the once variable

scale parameter element is now fixed. For  $B > 0$ , if  $\mathbb{K}$  is a compact set and the Lipschitz condition

$$\sup_{\mathbf{x} \in \mathbb{K}} |\log[k^p g(k\mathbf{x} - \mathbf{m}_1)] - \log[k^p g(k\mathbf{x} - \mathbf{m}_2)]| \leq B \|\mathbf{m}_1 - \mathbf{m}_2\|_1 \quad (2)$$

holds, then the following bound on the expected KL divergence for  $\hat{f}_{g,n,N}$  can be adapted from Li & Barron (1999).

**Theorem 22** (Li and Barron, 1999, Thm. 3). *Let  $g \in \mathcal{F}^3$  and suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_N$  is an IID random sample from a distribution with density  $f \in \mathcal{F}_{\mathbb{K},\beta}^5$ . For every  $\epsilon > 0$ , if (2) is satisfied and  $A = \bar{m} - \underline{m}$ , then under the restrictions of  $\mathcal{F}_{g,k}^8$ ,*

$$\mathbb{E}_f \left[ d_{\mathbb{K}}^{KL} \left( f, \hat{f}_{g,n,N} \right) \right] \leq \frac{\epsilon}{\beta} + \gamma^2 \frac{C^{*2}}{n} + \gamma \frac{2np}{N} \log(NAB\epsilon),$$

where  $\bar{f}^*$ ,  $\gamma$ , and  $C^*$  are as in defined in Theorem 19.

*Proof.* The original theorem provides the inequality

$$\mathbb{E}_f \left[ d_{\mathbb{K}}^{KL} \left( f, \hat{f}_{g,n,N} \right) \right] \leq d_{\mathbb{K}}^{KL} (f, \bar{f}^*) + \gamma^2 \frac{C^{*2}}{n} + \gamma \frac{2np}{N} \log(NAB\epsilon).$$

Using the same technique as in Corollary 21, we can bound the term  $d_{\mathbb{K}}^{KL} (f, \bar{f}^*)$  from above by  $d_{\mathbb{K}}^{KL} (f, \bar{f})$ , where upon for any  $\epsilon > 0$ , we can choose an  $\bar{f} \in \overline{\text{Conv}} \left( \mathcal{F}_{g,k}^8 \right)$  such that  $d_{\mathbb{K}}^{KL} (f, \bar{f}) < \epsilon/\beta$ .  $\square$

*Remark 23.* Since  $\epsilon$  can be made as small as we would like, the expected KL divergence between  $f$  and the MLE  $\hat{f}_{g,n,N}$  can be made arbitrarily small by choosing an increasing sequence of  $n$  that grows slower than  $N/\log N$  for  $N > N^*$ , for some  $N^*$ . For example, one can take  $n \propto \log N$ .

## 5 Concentration inequalities via results from Rakhlin et al. (2005)

We now proceed to utilize the theory of Rakhlin et al. (2005) to provide a concentration inequality for the MLE of finite mixtures of location-scale PDFs. Let  $\mathcal{N}(\Delta, \mathcal{F}, d)$  denote the  $\Delta$ -covering number of the class  $\mathcal{F}$ , with respect to the distance  $d$ . That is,  $\mathcal{N}(\Delta, \mathcal{F}, d)$  is the minimum number of  $\Delta$ -balls that is needed to cover  $\mathcal{F}$ , where a  $\Delta$ -ball around  $f$  (with centre not necessarily in  $\mathcal{F}$ ) is defined as  $\{g : d(f, g) < \delta\}$ ; see for example Kosorok (2008, Sec. 2.2.2). Further, define  $d_n$  as the empirical distance. That is for functions  $f$  and  $g$ , and realizations  $\mathbf{x}_1, \dots, \mathbf{x}_N$  of the random variables  $\mathbf{X}_1, \dots, \mathbf{X}_N$ , we have  $d_n^2(f, g) = N^{-1} \sum_{i=1}^N [f(\mathbf{x}_i) - g(\mathbf{x}_i)]^2$ . The following theorem can be

adapted from Rakhlin et al. (2005, Thm. 2.1).

**Theorem 24** (Rakhlin et al., 2005, Thm. 2.1). *Let  $g \in \mathcal{F}^3$  and suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_N$  is an IID random sample from a distribution with PDF  $f \in \mathcal{F}_{\mathbb{K}, \underline{\beta}}^5$  such that  $f(\mathbf{x}) < \bar{\beta}$  for all  $\mathbf{x} \in \mathbb{K}$ . If  $\hat{f}_{g,n,N}$  is the MLE for an  $n$ -component finite mixture of  $g$  (under the restrictions of  $\mathcal{F}_g^7$ ), then for any  $\epsilon > 0$*

$$\begin{aligned} & \mathbb{E}_f \left[ d_{\mathbb{K}}^{KL} \left( f, \hat{f}_{g,n,N} \right) \right] \\ & \leq \frac{\epsilon}{\underline{\beta}} + \frac{8\bar{\beta}^2}{n\underline{\beta}^2} \left( 2 + \log \frac{\bar{\beta}}{\underline{\beta}} \right) \\ & \quad + \frac{1}{\sqrt{N}} \left( \frac{\bar{\beta}C}{\underline{\beta}^2} \mathbb{E}_f \left( \int_0^{\bar{\beta}} \log^{1/2} \mathcal{N}(\Delta, \mathcal{F}_g^7, d_n) d\delta \right) + \frac{8\bar{\beta}}{\underline{\beta}} \right) \\ & \quad + \sqrt{\frac{t}{N}} \left( 4\sqrt{2} \log \frac{\bar{\beta}}{\underline{\beta}} \right), \end{aligned}$$

for some universal constant  $C$ , with probability at least  $1 - \exp(-t)$ .

*Proof.* The original statement of Rakhlin et al. (2005, Thm. 2.1) has  $d_{\mathbb{K}}^{KL}(f, \bar{f}^*)$  in place of  $\epsilon/\underline{\beta}$ . Thus, we obtain the desired result via the same technique as that used in Theorem 22.  $\square$

*Remark 25.* To make it directly comparable to Theorem 22, one can integrate out the probability statement of Theorem 24 to obtain the inequality in expectation

$$\begin{aligned} & \mathbb{E}_f \left[ d_{\mathbb{K}}^{KL} \left( f, \hat{f}_{g,n,N} \right) \right] \\ & \leq \frac{\epsilon}{\underline{\beta}} + \frac{8\bar{\beta}^2}{n\underline{\beta}^2} \left( 2 + \log \frac{\bar{\beta}}{\underline{\beta}} \right) \\ & \quad + \frac{1}{\sqrt{N}} \left[ \frac{\bar{\beta}C}{\underline{\beta}^2} \mathbb{E}_f \left( \int_0^{\bar{\beta}} \log^{1/2} \mathcal{N}(\delta, \mathcal{F}_g^7, d_n) d\delta \right) \right] \\ & \quad + \frac{1}{\sqrt{N}} \left( \frac{8\bar{\beta}}{\underline{\beta}} + 4\sqrt{2} \log \frac{\bar{\beta}}{\underline{\beta}} \right); \end{aligned}$$

see the proof of Rakhlin et al. (2005, Thm. 2.1) for details. The following corollary specializes the results of Theorem 24 to the hypothesis of Theorem 22.

**Corollary 26** (Rakhlin et al., 2005, Cor. 2.2). *Let  $g \in \mathcal{F}^3$  and suppose that  $\mathbf{X}_1, \dots, \mathbf{X}_N$  is an IID random sample from a distribution with density  $f \in \mathcal{F}_{\mathbb{K}, \underline{\beta}}^5$  such that  $f(\mathbf{x}) < \bar{\beta}$  for all  $\mathbf{x} \in \mathbb{K}$ . For every  $\epsilon > 0$ , if (2) is satisfied and  $A = \bar{m} - \underline{m}$ , under the restrictions of  $\mathcal{F}_{g,k}^8$ ,*

$$\mathbb{E}_f \left[ d_{\mathbb{K}}^{KL} \left( f, \hat{f}_{g,n,N} \right) \right] \leq \frac{\epsilon}{\underline{\beta}} + \frac{C_1}{n} + \frac{C_2}{\sqrt{N}},$$

where  $C_1$  and  $C_2$  are constants that depend on  $\underline{\beta}$ ,  $\bar{\beta}$ ,  $\gamma$ ,  $A$ ,  $B$ ,  $C^*$ , and  $p$ . Here,  $\gamma$  and  $C^*$  are as defined in Theorem 19.

*Remark 27.* Corollary 26 directly improves upon the result of Theorem 22 by allowing  $n$  and  $N$  to increase independently of one another and still be able to achieve an arbitrarily small bound on the expected KL divergence of the MLE for finite mixtures of location-scale PDFs, under the same hypothesis.

## 6 Truncated approximations

Thus far, we have discussed the approximation of compactly supported PDFs in  $\mathcal{F}_{\mathbb{K},\beta}^5$  by mixing distributions and finite mixture models constructed from PDFs, which are supported on  $\mathbb{R}^p$ . This is unappealing as it implies that although arbitrarily close, the PDF of any mixing distribution or any finite mixture model constructed via such PDFs will not integrate to unity over the support  $\mathbb{K}$ .

Thus, we are interested in the approximation a compactly supported PDF by PDFs on the same support. Without loss of generality, let  $\mathbb{K}_0$  be a compact subset of  $\mathbb{R}^p$ , such that  $\mathbf{0}$  is in an open subset of  $\mathbb{K}_0$ , where  $\mathbf{0}$  is the zero vector.

**Lemma 28.** *For any  $g \in \mathcal{F}^3$ ,  $\alpha_{\mathbb{K}_0,k}$  is an approximate identity on  $\mathbb{R}^p$  with  $k^* = \infty$ , where*

$$\alpha_{\mathbb{K}_0,k}(\mathbf{x}) = \begin{cases} k^p g(k\mathbf{x}) / \int_{\mathbb{K}_0} k^p g(k\mathbf{x}) d\mathbf{x}, & \text{if } \mathbf{x} \in \mathbb{K}_0, \\ 0, & \text{otherwise.} \end{cases}$$

*Proof.* Since  $g$  is a PDF,  $\alpha_{\mathbb{K}_0,k} \geq 0$  for all  $\mathbf{x} \in \mathbb{R}^p$  for any finite  $k$ , by definition, thus we have condition (i). By definition,

$$\int_{\mathbb{R}^p} \alpha_{\mathbb{K}_0,k}(\mathbf{y}) d\mathbf{y} = \frac{\int_{\mathbb{K}_0} k^p g(k\mathbf{y}) d\mathbf{y}}{\int_{\mathbb{K}_0} k^p g(k\mathbf{x}) d\mathbf{x}} = 1,$$

thus yielding condition (ii). Lastly, for any  $k$ ,  $\int_{\mathbb{K}_0} k^p g(k\mathbf{x}) d\mathbf{x} > 0$  since the set  $\mathbb{K}_0$  contains an open set and thus has non-zero Lebesgue measure. Since  $k^p g(k\mathbf{x}) \rightarrow \delta(\mathbf{0})$  as  $k \rightarrow \infty$ , we obtain condition (iii) and the desired result by noting that  $\delta(\mathbf{0}) / \lim_{k \rightarrow \infty} \int_{\mathbb{K}_0} k^p g(k\mathbf{x}) d\mathbf{x} = \delta(\mathbf{0})$  [cf. Prosperetti (2011, Tab. 2.1)].  $\square$

Define the class of truncated location-scale PDFs as

$$\mathcal{F}_{\mathbb{K}_0}^9 = \left\{ f : f(\mathbf{x}) = \frac{g(\mathbf{x})}{\int_{\mathbb{K}_0} g(\mathbf{x}) d\mathbf{x}}, \text{ for all } g \in \mathcal{F}^3 \right\},$$

for any compact set  $\mathbb{K}_0$  that satisfies the earlier mentioned conditions. We obtain the following direct analog to Corollary 6.

**Theorem 29.** *Let  $f$  be a PDF over  $\mathbb{R}^p$ . If  $g \in \mathcal{F}_{\mathbb{K}_0}^9$ , then for any  $\epsilon > 0$ , there exists a PDF  $f^*$  in*

$$\mathcal{F}_g^4 = \left\{ f^* : f^*(\mathbf{x}) = \int_{\mathbb{R}^p} k^p g(k\mathbf{x} - \mathbf{m}) f(\mathbf{m}) d\mathbf{m}, k \in \mathbb{N} \right\}$$

such that  $\|f - f^*\|_{\mathbb{R}^d, q} < \epsilon$ , for any  $q \in \mathbb{R}^+$ .

*Proof.* The proof is the same as that of Corollary 6, by making note of Remark 16. □

*Remark 30.* We note that although the result is obtained as bound of the  $\mathcal{L}_p$ -norm over  $\mathbb{R}^p$ , the result will also hold true for any compact subset  $\mathbb{K}$ . Thus, via application of Theorem 29, we can replace all occurrences of  $\mathcal{F}^3$  in the results from the previous sections by  $\mathcal{F}_{\mathbb{K}_0}^9$  (and also  $\mathbb{R}^p$  by  $\mathbb{K}_0$ , if we so wish).

**Corollary 31.** *Theorems 15, 18, 19, 22, and 24, and Corollaries 21 and 26 stay true when  $\mathcal{F}^3$  is replaced by  $\mathcal{F}_{\mathbb{K}_0}^9$ , for approximation of lower-bounded target PDFs over compact  $\mathbb{K}_0 \subset \mathbb{R}^p$ , where  $\mathbf{0}$  is in an open subset of  $\mathbb{K}_0$ .*

## References

- Barron, A. R. (1993). Universal approximation bound for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, IT-39, 930–945.
- Buhmann, M. D. (2003). *Radial Basis Functions: Theory And Implementation*. Cambridge University Press.
- Cheney, W. & Light, W. (2000). *A Course in Approximation Theory*. Pacific Grove: Brooks/Cole.
- DasGupta, A. (2008). *Asymptotic Theory Of Statistics And Probability*. New York: Springer.
- Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

- Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.
- Leroux, B. (2006). *Encyclopedia of Environmetrics*, volume 3, chapter Mixing Distribution. Wiley.
- Li, J. Q. & Barron, A. R. (1999). Mixture density estimation. In S. A. Solla, T. K. Leen, & K. R. Mueller (Eds.), *Advances in Neural Information Processing Systems*, volume 12 Cambridge: MIT Press.
- Light, W. A. (1993). Techniques for generating approximations via convolution kernels. *Numerical Algorithms*, 5, 247–261.
- Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*.
- Makarov, B. & Podkorytov, A. (2013). *Real Analysis: Measures, Integrals and Applications*. New York: Springer.
- McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.
- Prosperetti, A. (2011). *Advanced Mathematics for Applications*. Cambridge: Cambridge University Press.
- Rakhlin, A., Panchenko, D., & Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9, 220–229.
- Rossi, P. E. (2014). *Bayesian Non- and Semiparametric Methods and Applications*. Princeton: Princeton University Press.
- van de Geer, S. (2003). Asymptotic theory for maximum likelihood in nonparametric mixture models. *Computational Statistics and Data Analysis*, 41, 453–464.
- Walker, J. L. & Ben-Akiva, M. (2011). *A Handbook of Transport Economics*, chapter Advances in discrete choice: mixture models, (pp. 160–187). Edward Edgar.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- Willink, R. (2005). Normal moments and Hermite polynomials. *Statistics and Probability Letters*, 73, 271–275.
- Yona, G. (2011). *Introduction to Computational Proteomics*. Boca Raton: CRC Press.

Zeevi, A. J. & Meir, R. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Computation*, 10, 99–109.